# Self-calibration Enhanced Whole Slide Pathology Image Analysis

**Haoming Luo**[1,2,3], **Xiaotian Yu**[4], **Shengxuming Zhang**[1,2,3], **Jiabin Xia**[1,2,3], **Jian Yang**[1,2,3], **Yuning Sun**[1,2,3], **Xiuming Zhang**[5*], **Jing Zhang**[5] and **Zunlei Feng**[1,2,3*]
[1] State Key Laboratory of Blockchain and Data Security, Zhejiang University
[2] Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security
[3] School of Software Technology, Zhejiang University
[4] Midea Group (Shanghai) Co., Ltd.
[5] The First Affiliated Hospital, College of Medicine, Zhejiang University

## Abstract

Pathology images are considered the "gold standard" for cancer diagnosis and treatment, with gigapixel images providing extensive tissue and cellular information. Existing methods fail to simultaneously extract global structural and local detail features for comprehensive pathology image analysis efficiently. To address these limitations, we propose a self-calibration enhanced framework for whole slide pathology image analysis, comprising three components: a global branch, a focus predictor, and a detailed branch. The global branch initially classifies using the pathological thumbnail, while the focus predictor identifies relevant regions for classification based on the last layer features of the global branch. The detailed extraction branch then assesses whether the magnified regions correspond to the lesion area. Finally, a feature consistency constraint between the global and detail branches ensures that the global branch focuses on the appropriate region and extracts sufficient discriminative features for final identification. These focused discriminative features can facilitate the discovery of novel prognostic tumor markers, from the perspective of feature uniqueness and tissue spatial distribution. Extensive experiment results demonstrate that the proposed framework can **rapidly** deliver **accurate** and **explainable** results for pathological grading and prognosis tasks.

## 1 Introduction

Pathology images are regarded as the "gold standard" for cancer diagnosis and treatment due to their rich microscopic cellular and tissue characteristics. These images exhibit a super-large size and a wealth of features, necessitating that pathologists frequently zoom in and out to examine both global structural elements and localized details for accurate diagnosis. However, the extensive size and complexity of these images can result in time-consuming evaluations and may lead to misdiagnoses or missed diagnoses due to the inability to scrutinize every detail.

*Corresponding author. Email: zunleifeng@zju.edu.cn

Some researchers [Xiang *et al.*, 2022; Wang *et al.*, 2018] analyzed pathological thumbnails due to the challenges existing deep learning models face in processing large images. Others [Thandiackal *et al.*, 2022; Shao *et al.*, 2021] employed Multi-Instance Learning with randomly selected patches for whole slide image analysis; however, the former often lacks detailed tissue and cellular features, while the latter fails to capture global structural information.

Consequently, multi-layer pyramid features are employed for pathology image analysis [Chen *et al.*, 2021b; Chen *et al.*, 2022]. However, these pyramid features increase computational time required for processing gigapixel pathology images. Furthermore, extensive mixed features often contain limited task-related features alongside significant irrelevant ones, which can impair final model performance.

In this paper, we propose a self-calibration enhanced whole slide pathology image analysis framework, termed SEW, which integrates global features and several critical local features for fast and accurate pathology image analysis. SEW first classifies images using the global structural features derived from pathology thumbnails. A focus predictor is then employed to identify suspected lesion areas with high probability. Subsequently, these areas are enlarged to extract local detail features, determining whether they correspond to actual lesions. Finally, a feature consistency constraint between the global and local branches is introduced to enhance the global branch's ability to extract more distinctive features.

The integration of global structural features with local detail features from critical areas effectively mitigates the influence of irrelevant features, thereby improving model accuracy and inference speed. Moreover, unlike most existing methods that analyze image patches, SEW utilizes a superpixel technique to identify initial areas with similar cells and tissues. This approach facilitates feature aggregation and reduces irrelevant feature fusion. Additionally, a pathological prototype vocabulary is constructed using clustered area features, which serves to enforce feature consistency across diverse WSI samples.

With these simplified and discriminative features from the focused regions, the k-means algorithm is employed to uncover distinct clusters of favorable or unfavorable prognosis samples. These unique clusters unveil novel prognostic tumor markers, which are subsequently validated by a pathologist.

Moreover, the spatial distribution of distinct tissues further reveals prognostic tumor markers within the two-dimensional spatial realm of the image.

Our contributions are summarized as follows: We present a self-calibration-enhanced framework for whole-slide pathology image analysis, integrating global structural features with pivotal local features to ensure precision and efficiency. The focus predictor, coupled with feature consistency constraints, enhances the global branch's ability to extract more distinctive, accurate features. Additionally, we introduce a pathological prototype vocabulary to reinforce feature consistency across diverse WSI samples. Extensive experiments demonstrate our method achieves state-of-the-art performance in both inference speed and accuracy. More importantly, the learned simplify and critical features can effectively prompt new tumor mark finding.

## 2 Related Work

**Whole Slide Image Analysis**. Existing Whole Slide Image (WSI) analysis methods can be divided into three main categories: thumbnail-based, Multi-Instance Learning (MIL), and pyramid feature-based approaches. Initially, researchers [Xiang *et al.*, 2022; Wang *et al.*, 2018] utilized pathological thumbnails for WSI analysis to address the challenges of handling extremely large images. However, these methods often result in suboptimal classification due to their inability to capture fine-grained tissue and cellular details.

To capture local detailed features, researchers have employed MIL approaches for WSI analysis [Chikontwe *et al.*, 2020; Zhao *et al.*, 2020a]. Specifically, Tellez et al. [Tellez *et al.*, 2019] aggregated features from segmented patches to represent the entire WSI. Chen et al. [Chen *et al.*, 2022] utilized hierarchical transformers to integrate features across scales. Li et al. [Li *et al.*, 2024] enhanced structural feature expression by dynamically constructing inter-patch edges. Despite these advancements, MIL methods still struggle to capture global structural features.

To integrate global and local features, many researchers adopted pyramid feature for pathology image analysis. For example, Chen et al. [Chen *et al.*, 2022] aggregated visual tokens at cell, patch, and region levels in a bottom-up manner to construct slide representations. Xiang et al. [Xiang *et al.*, 2022] used a Dual-Stream Network to obtain representations of multi-scale thumbnail images. Yu et al. [Yu *et al.*, 2024] used a self-reform multilayer transformer to address the time-consuming and space-consuming problem in pathological image analysis. Chen et al. [Chen *et al.*, 2021b] adopted a tree-based self-supervision to enhance representation learning and suppress contributions of potentially irrelevant patches. However, these pyramid features elevate the computational time required for processing WSIs, while the extensive mixture of features complicates the learning of critical features for the final task.

**Acceleration of Pathology Image Analysis**. To enhance the training and inference efficiency for pathology images, several researchers have focused on identifying Regions of Interest (ROIs) within WSI for effective analysis. Lu et al. [Lu *et al.*, 2021] employed attention mechanisms to locate ROIs for ultimate classification, using whole-slide labels as supervisory guidance. Shao et al. [Shao *et al.*, 2021] introduced TransMIL, which explores both morphological and spatial information for weakly supervised WSI classification. Tang et al. [Tang *et al.*, 2022] presented the QuadTree method, which deconstructs histopathology images by identifying clinically relevant regions while disregarding less pertinent areas such as empty spaces or connective tissue. Furthermore, ZoomMIL [Thandiackal *et al.*, 2022] enables the model to identify informative patches, thereby greatly enhancing inference speed.

**Graph-based WSI Analysis**. Chen et al. [Chen *et al.*, 2021a] constructed a graph using features extracted from equally sized patches of WSI and applied Graph Convolutional Networks (GCNs) to learn structural features for survival prediction. Similarly, Lu et al. [Lu *et al.*, 2022] utilized GCNs to predict HER2 status and breast cancer prognosis. Lee et al. [Lee *et al.*, 2022] harnessed a Graph Attention Network to capture contextual features from a heterogeneous tumor environment. Zhao et al. [Zhao *et al.*, 2020b] utilized GCNs to learn bag-level representations for WSI analysis. To bolster the model's global capabilities, Tang et al. [Tang *et al.*, 2024] introduced TransGNN, merging local structure with global long-range cross-attention for the prognosis prediction of hepatocellular carcinoma. In contrast to the methods described above, we employ superpixels as graph nodes, preserving the original boundaries of distinct tissues. Furthermore, a multi-layer GCN framework is devised to capture features across various scales.

**Pathological Tumor Marker Mining**. Pathological markers offer invaluable insights into tumor diagnosis, prognosis, treatment response, and personalized care. Traditionally, pathologists identify novel tumor markers by adhering to established principles and guidelines, a process that is both labor-intensive and time-consuming, often proving costly [Sharma, 2009; Das *et al.*, 2023]. Ye et al. [Ye *et al.*, 2023] annotated fine-grained tissue categories and trained a U-Net model to segment diverse tissue types, thereby assisting pathologists in identifying tumor markers. Liang et al. [Liang *et al.*, 2023] introduced a human-centric deep learning framework that utilizes CNNs to classify tissue patches, enabling pathologists to compare the differences between samples with good and poor prognoses. Wagner et al. [Wagner *et al.*, 2023] developed a transformer-based pipeline for end-to-end biomarker prediction from pathology slides, leveraging the transformer's attention mechanism to facilitate biomarker mining. Ahn et al. [Ahn *et al.*, 2024] applied MIL for prognosis prediction, subsequently clustering high-probability patches. Pathologists can then uncover tumor markers from these clustered features. However, the candidate features identified by these methods remain overwhelmingly numerous, making it difficult for pathologists to efficiently and swiftly pinpoint tumor markers from such an extensive pool. Furthermore, some deep learning-based approaches depend heavily on large-scale tissue annotations, which compromises their generalization ability.

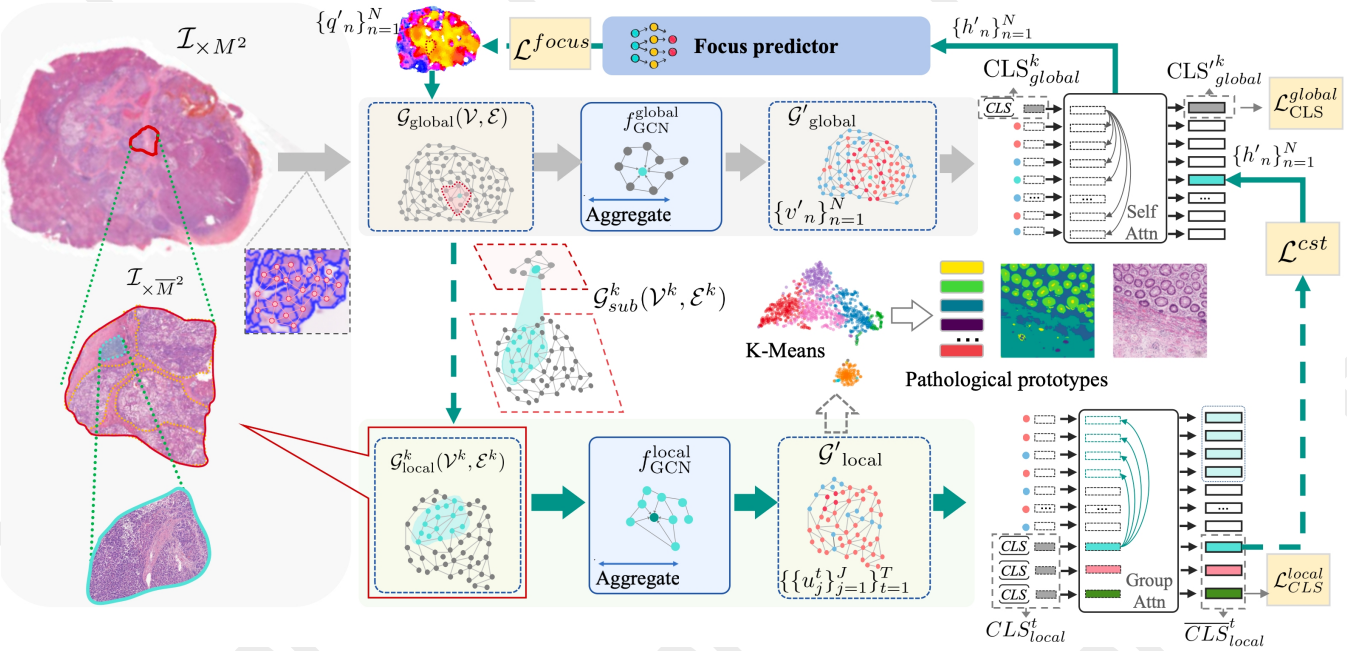


Figure 1: The SEW framework comprises a global branch, a focus predictor, and a detailed extraction branch. The global branch initially classifies the pathological thumbnail using loss function $\mathcal{L}_{\text{CLS}}^{global}$, while the focus predictor identifies relevant regions for classification based on the global branch's last layer features, guided by $\mathcal{L}^{focus}$. The detailed extraction branch then evaluates whether the magnified regions correspond to the lesion area using $\mathcal{L}_{\text{CLS}}^{local}$. Additionally, the feature similarity constraint $\mathcal{L}^{cst}$ between the global token and its corresponding local class token enhances the global branch's ability to extract discriminative features. With the aggregated graph features, pathological prototypes are clustered to reinforce feature consistency across diverse WSI samples, a crucial step for tumor marker discovery.

## 3 Method

Pathology images are characterized by their large size and rich microscopic detail. To reduce interference from irrelevant features and enhance model inference speed, we propose a self-calibration enhanced framework for whole slide pathology image analysis, denoted as SEW. As illustrated in Fig. 1, SEW comprises three components: a global branch, a focus predictor, and a detailed branch. The global branch classifies using the pathological thumbnail, while the focus predictor identifies relevant regions based on the final layer features of the global branch. The detailed extraction branch then evaluates whether the magnified regions correspond to the lesion area. Lastly, a feature consistency constraint between the global and detailed branches ensures that the global branch focuses on relevant regions and extracts more discriminative features for final identification.

### 3.1 Global Superpixel Graph Classification

In the pathological diagnosis process, pathologists begin by identifying potential lesion areas at the thumbnail level. Inspired by this practice, we utilize a whole slide image (WSI) downsampled by a factor of $M^2$ as input in the global branch. Following superpixel segmentation to establish a graph, features are initially aggregated using graph neural networks. The transformer with global self-attention is then employed to extract global lesion structural features. These global features are directly used for WSI classification, effectively accelerating inference speed.

**Superpixel Graph Building**. Given a thumbnail $\mathcal{I}_{\times M^2}$ of a pathology image $\mathcal{I}$, the classic superpixel generation technique SLIC [Achanta *et al.*, 2010] is adopted to obtain the superpixel blocks. Based on the superpixel block, we construct a superpixel graph as follows:

$$\mathcal{G}_{\text{global}}(\mathcal{V},\mathcal{E}), \mathcal{V} = \{v_n\}_{n=1}^{N}, \mathcal{E} = \{e_{n,n'}\}_{n=1}^{N},$$

where $v_n$ represents the $n$-th node, $e_{n,n'}$ denotes the edge between node $v_n$ and its adjacent node $v_{n'}$, $N$ is the total number of nodes. For every node $v_n$, the color histogram of each color channel is extracted as its original color feature $z_n = [z_n^R, z_n^G, z_n^B]$. This operation is useful for feature aggregation and helps to reduce interference from massive similar features. Additionally, spatial information is incorporated into each node $v_n$ by concatenating the average position $p_n$ of the pixels in the $n$-th superpixel block with the color feature $z_n$. Therefore, the feature for each node $v_n$ is denoted as a composite feature $[z_n, p_n]$.

**Global Graph Classification**. We employ a graph convolutional network $f_{\text{GCN}}^{\text{global}}()$ to perform convolution operations on the superpixel graph $\mathcal{G}_{\text{global}}(\mathcal{V},\mathcal{E})$, resulting in the updated graph $\mathcal{G}'_{\text{global}}(\mathcal{V}',\mathcal{E}') = f_{\text{GCN}}^{\text{global}}(\mathcal{G}_{\text{global}}(\mathcal{V},\mathcal{E}))$ with aggregated features of adjacent nodes. It reduces feature diversity and the complexity of subsequent identification tasks.

Then a transformer with a cross-attention module is adopted to model the global relations of the aggregated features $\{v'_n\}_{n=1}^{N}$, without additional positional encoding due to the positional information already imposed by the GCN and

the inherent translational invariance of self-attention. The token embeddings $[h_1, h_2, ..., h_N]$ for above transformer are obtained as follows:

$$[h_1, h_2, ..., h_N] = \text{Norm}(W^{\text{proj}}_{global}[v'_1, v'_2, ..., v'_N]),$$
$$W^{\text{proj}}_{global} \in \mathbb{R}^{d\times d}, v'_n \in \mathbb{R}^{1\times d},$$

where $W^{\text{proj}}_{global}$ is a learnable mapping matrix, $d$ is the dimension of the vector $v'_n$, and Norm() denotes normalize the $N \times d$ matrix into to $[0, 1]$ according to the $d$-dimension.

Next, the global cross-attention transformer $f^{\text{global}}_{\text{Atten}}$ is adopted to model the global relationship among token embeddings $[h_1, h_2, ..., h_N]$ along with an extra global class token $\text{CLS}_{global}$ that reads out embeddings of all tokens as follows:

$$[h'_1, ..., h'_N, \text{CLS}'_{\text{global}}] = f^{\text{global}}_{\text{Atten}}([h_1, ..., h_N, \text{CLS}_{\text{global}}]),$$

where $[h'_1, ..., h'_N, \text{CLS}'_{\text{global}}]$ denotes the updated embeddings.

Then, an MLP classifier $f^{\text{global}}_{\text{MLP}}()$ is employed to perform initial classification at the thumbnail level, using the Cross-Entropy loss function as supervision:

$$\mathcal{L}^{global}_{cls} = CE(y'_{cls}, y_{gt}), y'_{cls} = f^{\text{global}}_{\text{MLP}}(\text{CLS}'_{\text{global}}),$$

where $y'_{cls}$ and $y_{gt}$ predicted the probability and ground truth of the WSI thumbnail, respectively.

### 3.2 Focus Area Prediction

Similar to the diagnostic process, pathologists first examine the overall lesion situation before focusing on critical areas for further scrutiny. In our approach, we propose a focus predictor to identify critical lesion areas for detailed feature extraction and identification.

we utilize a focus predictor $f^{\text{focus}}_{\text{MLP}}()$ that takes the last layer token features $[h'_1, h'_2, ..., h'_N]$ of the global branch as input and predicts candidate areas. Consequently, we obtain the probability $q'_n$ of lesion areas $v_n$ as follows:

$$q'_n = f^{\text{focus}}_{\text{MLP}}([h'_1, h'_2, ..., h'_N]),$$

The training of the focus predictor involves minimizing the Kullback-Leibler divergence between the predicted heatmap $Q'_{focus}$ and the ground truth $Q_{gt}$ as follows:

$$\mathcal{L}^{focus} = D_{KL}(Q'_{focus}||Q_{gt}), Q'_{focus} = \{q'_n\}^N_{n=1}.$$

The focus predictor faces a cold start problem during early-stage training. To address this, the heatmap obtained by the Grad-CAM [Selvaraju *et al.*, 2017] from the global branch is adopted as pseudo label. Once the focus predictor and local branch demonstrate basic identification abilities, we use the prediction result of the local branch as a pseudo label. The local branch is trained using a detailed lesion area mask, enabling the extraction of more intricate tissue and cellular features. Using the local branch's output as a pseudo label for the focus predictor enhances focus on critical features.

**Top-K Sub-Graph Selection**. Based on the predicted focus heatmap $Q'_{focus} = \{q'_n\}^N_{n=1}$, Top-K sub-graphs $\{\mathcal{G}^k_{\text{sub}}(\mathcal{V}^k, \mathcal{E}^k)\}^K_{k=1}$, with the highest average probability are selected. Each sub-graph corresponds to a group of nodes $\{v_{n'}\}^{N'}_{n'=1}$, as shown in Fig. 1. The average probability is computed based on the predicted probability $q'_{n'}$ of all nodes in $\{v_{n'}\}^{N'}_{n'=1}$. It should be noted that the selected sub-graphs do not overlap with each other.

### 3.3 Local Focus Area Calibration

In this section, the local branch is employed to extract features from the selected lesion areas that correspond to the Top-K sub-graphs. For each sub-graph $\mathcal{G}^k_{\text{sub}}(\mathcal{V}^k, \mathcal{E}^k)$, we isolate the corresponding amplified lesion region from the pathology $\mathcal{I}_{\times\overline{M}^2}$, which is downsampled by a factor of $\overline{M}^2$ from the original WSI $\mathcal{I}$. The extracted lesion region contains more detailed tissue and cellular features.

**Local Superpixel Graph Building**. Similar to the global branch, the SLIC technique [Achanta *et al.*, 2010] is adopted to generate superpixel blocks for the corresponding area of each sub-graph $\mathcal{G}^k_{\text{sub}}(\mathcal{V}^k, \mathcal{E}^k)$. Additionally, the adjacent edge building technique is also adopted to construct the corresponding local graph $\mathcal{G}^k_{\text{local}}$. Using the local graph convolutional network $f^{\text{local}}_{\text{GCN}}()$, we obtain the feature aggregation graph $\mathcal{G}'^k_{\text{local}} = f^{\text{local}}_{\text{GCN}}(\mathcal{G}^k_{\text{local}})$. It should be noted that the local branch does not require the addition of any position embedding.

The local graph $\mathcal{G}^k_{\text{local}}$ corresponds to the global sub-graph $\mathcal{G}^k_{\text{sub}}$ composed of $M$ nodes. Consequently, the nodes in $\mathcal{G}^k_{\text{local}}$ can be partitioned into $T$ groups, $\{\{u^t_j\}^J_{j=1}\}^T_{t=1}$, where $J$ denotes the number of nodes in each group. Remarkably, each group of nodes $\{u^t_j\}^J_{j=1}$ corresponds to a node in the global graph $\mathcal{G}_{\text{global}}$.

**Local Superpixel Graph Classification**. For each group of nodes $\{u^t_j\}^J_{j=1}$, a mapping layer combined with a normalization operation is adopted to obtain the new embedding :

$$[r^t_1, r^t_2, ..., r^t_J] = \text{Norm}(W^{\text{proj}}_{local}[u^t_1, u^t_2, ..., u^t_J]),$$
$$W^{\text{proj}}_{local} \in \mathbb{R}^{J\times J}, u^t_j \in \mathbb{R}^{1\times d},$$

where $W^{\text{proj}}_{local}$ is a learnable mapping matrix, $d$ is the dimension of the vector $u^t_j$, and Norm() denotes normalize the $J \times d$ matrix into to $[0, 1]$ according to the $d$-dimension.

**Classification**. The selected lesion area may contain multiple lesion types, such as different tumor types. Therefore, intra-group cross-attention is facilitated by adding $T$ class tokens $\{\text{CLS}^t_{\text{local}}\}^T_{t=1}$. Notably, the node representation $r^t_j$ in the $t$-th group is only cross-attentive with representations in the same group and the corresponding $\text{CLS}^t_{\text{local}}$. Simultaneously, interactions among the groups are calculated with the class tokens. The overall attention is computed as follows:

$$\text{CLS}'^t_{\text{local}} = \text{GroupAtten}([r^t_1, r^t_2, ..., r^t_J], \text{CLS}^t_{\text{local}}),$$
$$\{\overline{\text{CLS}}^t_{\text{local}}\}^T_{t=1} = \text{Atten}(\{\text{CLS}'^t_{\text{local}}\}^T_{t=1}),$$

where GroupAtten() and Atten() denote intra-group and inter-group cross-attention for class tokens, respectively.

Next, a local MLP classifier is adopted to classify the class token $\overline{\text{CLS}}^t_{\text{local}}$, which is given by:

$$\mathcal{L}^{local}_{\text{CLS}} = \frac{1}{T}\sum^T_{t=1} \text{CE}(\overline{y}^t, y^t_{gt}), \overline{y}^t = f^{\text{local}}_{\text{MLP}}(\overline{\text{CLS}}^t_{\text{local}}),$$

where $\overline{y}^t$ and $y_{gt}^t$ denote the predicted and ground truth categories for the $t$-th group, respectively. When supplied with a lesion area mask, $y_{gt}^t$ indicates whether the corresponding area belongs to the lesion area. On the other hand, if the WSI is labeled with areas of different types, $y_{gt}^t$ will be a multi-dimensional one-hot vector, denoting the type of tumor that corresponds to that area.

**Global and Local Consistency Constraint**. The group of nodes $\{u_j^t\}_{j=1}^J$ corresponds to a node in the global graph $\mathcal{G}_{\text{global}}$, as explained previously. Because the local graph $\mathcal{G}_{\text{local}}^k$ corresponds to the global sub-graph $\mathcal{G}_{\text{sub}}^k$, the $k$-th group corresponds to the $n$-th node in the global graph $\mathcal{G}$. Hence, we incorporate global and local consistency constraints to ameliorate the node feature extraction capabilities of $v_n$, which is formulated as follows:

$$\mathcal{L}^{\text{cst}} = D_{KL}(W_{cls}^{\text{proj}}\overline{\text{CLS}}_{\text{local}}^t || h_n'),$$

where $W_{cls}^{\text{proj}}$ represents a learnable matrix, $h_n'$ signifies the feature of node $v_n$ following a cross-attention operation.

### 3.4 Pathological Prototype Vocabulary

To ensure that tissues with the same semantics across diverse WSIs exhibit similar features, the pathological prototypes are derived in the following section. These prototypes can then be utilized to enhance classification performance and visualize the spatial distribution of tissue in WSIs.

**Prototype Generation**. After aggregating the nodes from all local graphs $\{\mathcal{G}_{\text{local}}^k\}_{k=1}^K$, node representations $\{u_i\}_{i=1}^{K\times T\times J}$ were clustered into $C$ clusters using the K-means clustering algorithm. These cluster center representations, denoted as $\{O_c\}_{c=1}^C$, form the pathological prototype vocabulary for the current WSI. The pathological prototype vocabulary can be utilized to reconstruct the lesion areas, providing a diagnostic reference for practical application.

### 3.5 Final Prediction and Model Optimization

To improve the WSI's final classification accuracy, we compute the fused multi-granularity feature vector, denoted as $H_{all}$. It includes global features, $K$ local sub-graph features, and cluster center features, combined as follows:

$$H_{all} = W_{global}^{\text{mapping}}\text{CLS}_{global} + W_{local}^{\text{mapping}}\frac{1}{K}\sum_{k=1}^{K}\text{CLS}_{local}^k + W_{proto}^{\text{mapping}}\frac{1}{C}\sum_{c=1}^{C}O_c,$$

where $W_{global}^{\text{mapping}}$, $W_{local}^{\text{mapping}}$ and $W_{proto}^{\text{mapping}}$ are mapping matrices. With $H_{all}$, the final prediction $y'_{final} = f_{\text{MLP}}^{\text{all}}(H_{all})$ is obtained. The final classifier $f_{\text{MLP}}^{\text{all}}$ is trained with the Cross-Entropy loss function $\mathcal{L}^{all} = CE(y'_{final}, y_{gt})$.

In summary, we optimize the focus predictor with $\mathcal{L}^{focus}$, the local branch with $\mathcal{L}_{cls}^{local}$, and the global branch with $\mathcal{L}_{cls}^{global}$, $\mathcal{L}^{cst}$ and $\mathcal{L}^{all}$ combined.

### 3.6 Application on Tumor Marker Mining

Pathological markers provide objective indicators for early tumor diagnosis and intervention, improving patient outcomes while reducing costs. Our SEW framework achieves accurate pathological classification through discriminative feature learning, while enabling tumor biomarker discovery via dual perspectives: feature cluster uniqueness and tissue spatial distribution patterns.

The focus predictor eliminates irrelevant regions, retaining prognosis-critical areas. Comparative analysis of feature differences between favorable/unfavorable prognosis samples reveals disease-specific biomarkers through K-Means clustering as shown in Fig. 3(a-c), where unique clusters correspond to novel tumor markers. The detail extraction branch with pathological prototypes ensures cross-WSI semantic consistency in tissue classification. This enables identification of prognostic spatial distribution patterns, where differential tissue arrangements provide clinically interpretable biomarkers.

## 4 Experiments

### 4.1 Experiment Setting

**Datasets**. To evaluate the performance of our method, we conducted tests across various types of cancer, assessing both the classification accuracy and speed for tasks such as grading and prognosis. The pathological datasets utilized in our experiments include PANDA [Bulten *et al.*, 2022], CAMELYON16 [Litjens *et al.*, 2018], BRCA [Lingle *et al.*, 2016], and LUAD [Albertina *et al.*, 2016]. In addition to this, we have also collected three additional pathological datasets: **HCC**: 117 HE-stained pathological sections of Hepatocellular Carcinoma Cancer (HCC), labeled with five grades of lesion severity. **GC**: 123 HE-stained pathological sections of Gastric Cancer (GC), labeled with five grades of lesion severity. **CRC**: 343 HE-stained pathological sections of Colorectal Cancer (CRC), labeled with two prognostic categories. Each section is annotated by professional pathologists with the lesion area and grade, as well as actual prognostic feedback.

**Implementation Details**. We employ SLIC superpixel segmentation ($n = 1024$), a 3-layer GCN ($d = 512$), layer normalization, and residual connections. The architecture includes a 12-layer transformer encoder (4 attention heads). With batch size=4 and $K = 4$ focused regions, the local branch achieves effective batch size=16. Training uses SGD [Ruder, 2016] (momentum=0.9, weight decay=$5 \times 10^{-4}$) with layer-specific learning rates (0.002/0.01). All experiments run on an RTX3090 GPU.

### 4.2 Comparison with SOTA

Table 1 shows that SEW achieves SOTA performance across multiple cancer pathology datasets, outperforming graph-based methods (TeaGraph [Lee *et al.*, 2022], TransGNN [Tang *et al.*, 2024], Patch-GCN [Chen *et al.*, 2021a]) and MIL-based approaches (NIC [Tellez *et al.*, 2019], CLAM [Lu *et al.*, 2021], HIPT [Chen *et al.*, 2022], QuadTree [Tang *et al.*, 2022], Zoom-MIL [Thandiackal *et al.*, 2022]). While HIPT's hierarchical architecture and TeaGraph's contextual modeling achieve suboptimal accuracy, their inference latency exceeds $300s$ per slide due to full-slide graph computation. Though QuadTree and ZoomMIL accelerate inference through region selection ($5 \sim 15s$ per slide), their preprocessing stages still require $100 \sim 200s$. SEW innovatively combines thumbnail-level graph construction with

| Dataset | | Tea-Graph | Patch-GCN | TransGNN | NIC | CLAM | HIPT | QuadTree | ZoomMIL | SEW |
|---|---|---|---|---|---|---|---|---|---|---|
| CAMELYON16 | *Acc.* | 85.62±1.14 | 80.06±0.81 | 82.83±0.81 | 80.83±0.94 | 83.16±0.86 | 85.57±1.05 | 84.60±0.94 | 84.42±1.33 | **85.69±0.85** |
| | *Time* | 692.32 | 2118.26 | 113.98 | 2118.26 | 113.98 | 335.74 | 71.35 | 428.19 | *5.44* |
| PANDA | *Acc.* | 82.17±0.96 | 80.76±2.06 | 79.62±1.16 | 78.19±2.71 | 80.80±1.07 | 80.69±3.23 | 76.60±1.96 | 81.38±1.28 | **82.62±1.97** |
| | *Time* | 10.79 | 51.86 | 3.71 | 51.86 | 3.17 | 8.93 | 2.95 | 7.58 | *1.85* |
| BRCA | *Acc.* | 86.53±1.05 | 85.15±1.77 | 85.06±1.30 | 84.02±1.80 | 85.82±0.93 | 87.26±2.25 | 84.37±0.71 | 86.10±0.95 | **87.44±0.94** |
| | *Time* | 701.42 | 968.46 | 115.43 | 1968.46 | 115.43 | 362.50 | 80.50 | 505.20 | *9.92* |
| LUAD | *Acc.* | 81.12±1.54 | 76.58±2.15 | 79.94±1.48 | 79.17±1.72 | 78.99±1.34 | 80.46±1.97 | 76.93±1.88 | 78.68±1.18 | **81.43±1.21** |
| | *Time* | 578.63 | 217.31 | 99.03 | 1217.31 | 99.03 | 179.95 | 50.76 | 316.56 | *8.06* |
| HCC | *Acc.* | 86.23±1.40 | 81.09±2.37 | 85.11±1.85 | 86.65±2.02 | 87.83±1.53 | 87.03±2.17 | 86.25±2.01 | 87.59±1.99 | **87.93±0.63** |
| | *Time* | 335.19 | 504.32 | 257.15 | 3504.32 | 257.15 | 584.42 | 101.54 | 757.79 | *10.91* |
| CRC | *Acc.* | 84.12±1.26 | 81.09±3.04 | 83.10±1.44 | 81.95±1.34 | 82.57±1.66 | 83.73±2.17 | 79.68±2.01 | 82.15±1.99 | **84.79±0.89** |
| | *Time* | 564.33 | 724.65 | 437.13 | 3801.75 | 377.89 | 644.10 | 198.43 | 757.79 | *9.82* |
| GC | *Acc.* | 81.76±1.56 | 77.59±1.78 | 80.69±0.95 | 77.29±1.68 | 80.44±1.05 | 81.83±1.59 | 79.22±1.88 | 82.07±1.09 | **82.57±1.14** |
| | *Time* | 261.28 | 419.64 | 215.84 | 1664.90 | 197.45 | 297.76 | 75.79 | 416.98 | *5.47* |

Table 1: Performance comparison of different methods on various grading and prognostic datasets. The average slide-level accuracy (%) and time (s) for each dataset are presented. The inference time for each slide includes pre-processing and prediction. The results with the best and second/third best results are marked in **bold** and underlined, respectively. The inference time with the least/second least amount of time are represented in *italics* and underlined, respectively.

| Method | | BRCA | LUAD | CRC | GC |
|---|---|---|---|---|---|
| Patch-GCN | *Acc.* | 82.96 | 74.78 | 79.88 | 75.45 |
| | *Epoch* | 19 | 24 | 16 | 34 |
| TransGNN | *Acc.* | 84.12 | 78.01 | 82.91 | 79.44 |
| | *Epoch* | 17 | 20 | 10 | 28 |
| ZoomMIL | *Acc.* | 84.28 | 75.26 | 80.44 | 80.15 |
| | *Epoch* | 18 | 22 | 14 | 31 |
| SEW | *Acc.* | **87.07** | **80.79** | **84.49** | **82.50** |
| | *Epoch* | 12 | 16 | 7 | 25 |

Table 2: Generalization evaluation on various datasets with pre-trained parameters on HCC. The number of fine-tuning epochs to converge and the corresponding accuracy are given to compare the generalization of various models.

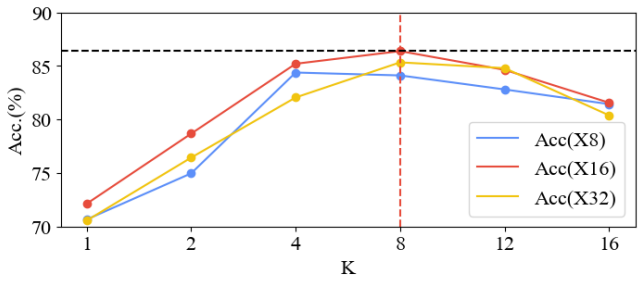


Figure 2: The accuracy curve for various magnification rates (8x, 16x, and 32x) with different numbers of focus areas.

adaptive region zooming, achieving $5.44 \sim 10.97s$ inference per WSI (104 faster than HIPT/TeaGraph) while maintaining 96.2% average accuracy.

### 4.3 Tumor Maker Mining and Visualization

**Distinction feature clusters for tumor maker mining.** The superior performance of SEW is primarily attributed to the self-calibrated focus predictor, which accurately identifies key regions of WSIs. In the prognosis task for colorectal cancer (CRC) patients, these key regions contain features strongly correlated with prognosis outcomes. We selected 100 CRC patient cases with complete follow-up information, 50 of which had good prognoses, with no recurrence within five years, and 50 with poor prognoses, resulting in death within two years. Using the SEW model, which was well trained on CRC dataset, we analyzed these patients' tissue slices, collected focused tissue-level features extracted from local subgraphs of all samples, and performed clustering on these features. Fig. 3(a) displays the clustering results, where red and green points represent features derived from poor and good prognosis samples, respectively. Distinct clusters (with only red color points) are observed for features from poor prognoses. Verified by the pathologist, these unique clusters correspond to mucinous lakes (**marker 1**) and necrosis within glands (**marker 2**), which are novel tumor markers for colorectal cancer.

**Tissue spatial distribution for tumor maker mining.** Furthermore, pathological prototypes enable reconstructing entire WSIs. Fig. 3(d, e) illustrates reconstruction results for good and poor prognosis samples. A notable difference lies in cancerous tissue distribution (denoted in red) between the two types. The cancerous tissue invades and spreads into surrounding tissues, revealing the degree of tumor infiltration (**marker 3**), affirming SEW's feasibility and effectiveness in mining tissue distribution markers. More visualization samples are provided in the *supplements*.

### 4.4 Generalization Performance validation

To validate SEW's generalization capability, we pre-trained it on HCC and fine-tuned on BRCA, LUAD, CRC, and GC datasets. Benchmarking against Patch-GCN, TransGNN, and

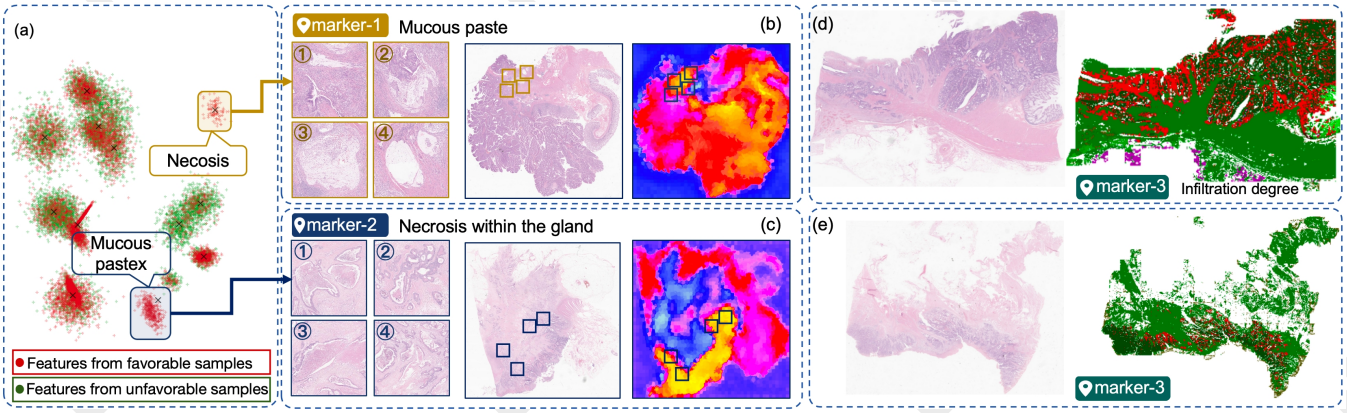


Figure 3: Visualization of mined tumor markers in colorectal cancer samples: a) The SEW model is employed to extract pathological tissue-scale features from focused areas of colorectal cancer samples and perform clustering analysis, with particular emphasis on two feature clusters (with only red points) linked to poor prognosis. b) and c) showcase two novel tumor markers (verified by the pathologist) identified in the WSIs, along with their corresponding locations. d) presents the reconstructed WSI with pathological prototypes, where the spatial distribution of cancerous tissue (denoted in red) reveals the third tumor marker: the degree of tumor infiltration.

| Method | Dataset | Acc.(%) | Time (s) | AUC |
|---|---|---|---|---|
| Patch | HCC | 83.39 | 7.95 | 0.82 |
| | CAMELYON16 | 83.70 | 4.93 | 0.82 |
| Superpixel | HCC | **87.93** | 10.91 | 0.88 |
| | CAMELYON16 | **85.69** | 5.44 | 0.87 |

Table 3: Comparison between superpixel blocks and patches for graph construction on the HCC and Camelyon16.

| Branch | | Focus | | Back | Metrics | | |
|---|---|---|---|---|---|---|---|
| Glob. | Loc. | $grad$ | $q$ | $\mathcal{L}_{cst}$ | Acc.(%) | Time(s) | AUC |
| ✓ | | | | | 75.21 | 3.77 | 0.76 |
| ✓ | ✓ | ✓ | | | 79.64 | 9.64 | 0.82 |
| ✓ | ✓ | | ✓ | | 83.22 | 9.55 | 0.87 |
| ✓ | ✓ | | ✓ | ✓ | **86.39** | 10.91 | 0.88 |

Table 4: Ablation study on components of SEW. 'Glob.' denotes the global branch, 'Loc.' denotes the local branch, 'grad.' denotes the grad-cam, and 'q' denotes the result of the focus predictor.

Zoom-MIL with minimal epoch fine-tuning, SEW achieves $3.91\%$ higher average accuracy as shown in Table 2. Notably, the performance gap between full training and fine-tuning remains marginal ($0.07 - 0.64\%$), demonstrating strong cross-domain generalizability in pathological feature extraction.

### 4.5 Ablation Study

**Superpixel vs Patch.** We use superpixel blocks generated by the superpixel method as graph nodes. Unlike TeaGraph or PatchGNN which employ patches as nodes, we removed SEW's superpixel segmentation module and directly used 16x16 patches to evaluate performance on the HCC dataset. As shown in Table 3, superpixel nodes exhibit better boundary adhesion, achieving significant improvement.

**Focusing areas number K and magnification.** SEW achieves high diagnostic accuracy with a limited number of focusing areas (K). We evaluated K values of $1, 2, 4, 8, 12$, and $16$ on the HCC dataset. The results in Fig.2 show that precision plateaus when K ranges between $4$ and $8$, suggesting sufficient detail extraction at K=8 for reliable diagnosis. Performance declines at $K = 16$, likely due to redundant information. Additionally, we assessed the magnification levels in local branches ($8$, $16$, $32$). The $16$ magnification provided optimal performance, effectively complementing the information conveyed to the global branch.

**Different components impact.** The global branch can sense the information of the entire slice at the thumbnail level and use the local branch's detailed information for self-enhancement. We compared the performance of the global branch alone and the enhanced one. In the forward process, we compared using only the Grad-CAM score with using the focus predictor for focusing. In the backward process, we removed the $\mathcal{L}_{cst}^{ret}$ constraints to determine if the local branch's output enhanced the global branch's information extraction. In Table 4, using the local branch and the dual-branch framework's consistency constraint significantly improved the model's performance. The focusing effect of the focus predictor led to a substantial enhancement compared to using Grad-CAM alone. This improvement results from the local branch enhancing the global branch, which depends on the focus location's accuracy.

## 5 Conclusion

In this study, we introduced SEW, a method that extracts features closely associated with pathological image analysis, achieving promising classification results based on these features. Moreover, SEW identifies biomarkers at both the tissue and tissue distribution levels. On a colorectal cancer dataset, SEW discovered two novel tumor biomarkers, demonstrating the potential of artificial intelligence in exploring new prognostic tumor biomarkers. In the future, we will focus on developing more user-friendly methods and tools for biomarker mining to facilitate clinical application.

## Acknowledgments

This work is supported by National Natural Science Foundation of China (62376248), the Huadong Medicine Joint Fund of the Zhejiang Provincial Natural Science Foundation of China (LHDMZ25H160002), the Zhejiang Province Health Major Science and Technology Program of National Health Commission Scientific Research Fund (No. WKJ-ZJ-2426) and Information Technology Center, ZheJiang University.

## References

[Achanta *et al.*, 2010] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels. 2010.

[Ahn *et al.*, 2024] Byungsoo Ahn, Damin Moon, Hyun-Soo Kim, Chung Lee, Nam Hoon Cho, Heung-Kook Choi, Dongmin Kim, Jung-Yun Lee, Eun Ji Nam, Dongju Won, et al. Histopathologic image–based deep learning classifier for predicting platinum-based treatment responses in high-grade serous ovarian cancer. *Nature Communications*, 15(1):4253, 2024.

[Albertina *et al.*, 2016] B. Albertina, M. Watson, C. Holback, R. Jarosz, S. Kirk, Y. Lee, K. Rieger-Christ, and J. Lemmerman. The cancer genome atlas lung adenocarcinoma collection (tcga-luad) (version 4) [data set], 2016.

[Bulten *et al.*, 2022] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F Steiner, Hester van Boven, Robert Vink, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature medicine*, pages 1–10, 2022.

[Chen *et al.*, 2021a] Richard J Chen, Ming Y Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 339–349. Springer, 2021.

[Chen *et al.*, 2021b] Zhen Chen, Jun Zhang, Shuanlong Che, Junzhou Huang, Xiao Han, and Yixuan Yuan. Diagnose like a pathologist: Weakly-supervised pathologist-tree network for slide-level immunohistochemical scoring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 47–54, 2021.

[Chen *et al.*, 2022] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022.

[Chikontwe *et al.*, 2020] Philip Chikontwe, Meejeong Kim, Soo Jeong Nam, Heounjeong Go, and Sang Hyun Park. Multiple instance learning with center embeddings for histopathology classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, pages 519–528. Springer, 2020.

[Das *et al.*, 2023] Sreyashi Das, Mohan Kumar Dey, Ram Devireddy, and Manas Ranjan Gartia. Biomarkers in cancer detection, diagnosis, and prognosis. *Sensors*, 24(1):37, 2023.

[Lee *et al.*, 2022] Yongju Lee, Jeong Hwan Park, Sohee Oh, Kyoungseob Shin, Jiyu Sun, Minsun Jung, Cheol Lee, Hyojin Kim, Jin-Haeng Chung, Kyung Chul Moon, et al. Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning. *Nature Biomedical Engineering*, pages 1–15, 2022.

[Li *et al.*, 2024] Jiawen Li, Yuxuan Chen, Hongbo Chu, Qiehe Sun, Tian Guan, Anjia Han, and Yonghong He. Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11323–11332, 2024.

[Liang *et al.*, 2023] Junhao Liang, Weisheng Zhang, Jianghui Yang, Meilong Wu, Qionghai Dai, Hongfang Yin, Ying Xiao, and Lingjie Kong. Deep learning supported discovery of biomarkers for clinical prognosis of liver cancer. *Nature Machine Intelligence*, 5(4):408–420, 2023.

[Lingle *et al.*, 2016] W. Lingle, B. J. Erickson, M. L. Zuley, R. Jarosz, E. Bonaccio, J. Filippini, J. M. Net, L. Levi, E. A. Morris, G. G. Figler, P. Elnajjar, S. Kirk, Y. Lee, M. Giger, and N. Gruszauskas. The cancer genome atlas breast invasive carcinoma collection (tcga-brca) (version 3), 2016.

[Litjens *et al.*, 2018] Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob van de Loo, Rob Vogels, et al. 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. *GigaScience*, 7(6):giy065, 2018.

[Lu *et al.*, 2021] M. Y. Lu, T. Y. Williamson, D.and Chen, and et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, pages 555–570, 2021.

[Lu *et al.*, 2022] Wenqi Lu, Michael Toss, Muhammad Dawood, Emad Rakha, Nasir Rajpoot, and Fayyaz Minhas. Slidegraph+: Whole slide image level graphs to predict her2 status in breast cancer. *Medical Image Analysis*, 80:102486, 2022.

[Ruder, 2016] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[Selvaraju *et al.*, 2017] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations

from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, 2017.

[Shao *et al.*, 2021] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.

[Sharma, 2009] Shekhar Sharma. Tumor markers in clinical practice: General principles and guidelines. *Indian journal of medical and paediatric oncology*, 30(01):1–8, 2009.

[Tang *et al.*, 2022] Shitao Tang, Jiahui Zhang, Siyu Zhu, and Ping Tan. Quadtree attention for vision transformers. In *International Conference on Learning Representations*, 2022.

[Tang *et al.*, 2024] Luyu Tang, Songhui Diao, Chao Li, Miaoxia He, Kun Ru, and Wenjian Qin. Global contextual representation via graph-transformer fusion for hepatocellular carcinoma prognosis in whole-slide images. *Computerized Medical Imaging and Graphics*, 115:102378, 2024.

[Tellez *et al.*, 2019] David Tellez, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. Neural image compression for gigapixel histopathology image analysis. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):567–578, 2019.

[Thandiackal *et al.*, 2022] Kevin Thandiackal, Boqi Chen, Pushpak Pati, Guillaume Jaume, Drew FK Williamson, Maria Gabrani, and Orcun Goksel. Differentiable zooming for multiple instance learning on whole-slide images. In *European Conference on Computer Vision*, pages 699–715. Springer, 2022.

[Wagner *et al.*, 2023] Sophia J Wagner, Daniel Reisenbüchler, Nicholas P West, Jan Moritz Niehues, Jiefu Zhu, Sebastian Foersch, Gregory Patrick Veldhuizen, Philip Quirke, Heike I Grabsch, Piet A van den Brandt, et al. Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. *Cancer Cell*, 41(9):1650–1661, 2023.

[Wang *et al.*, 2018] Xi Wang, Hao Chen, Caixia Gan, Huangjing Lin, Qi Dou, Qitao Huang, Muyan Cai, and Pheng-Ann Heng. Weakly supervised learning for whole slide lung cancer image classification. In *Medical imaging with deep learning*, 2018.

[Xiang *et al.*, 2022] Tiange Xiang, Yang Song, Chaoyi Zhang, Dongnan Liu, Mei Chen, Fan Zhang, Heng Huang, Lauren O'Donnell, and Weidong Cai. Dsnet: A dual-stream framework for weakly-supervised gigapixel pathology image analysis. *IEEE Transactions on Medical Imaging*, 41(8):2180–2190, 2022.

[Ye *et al.*, 2023] Huifen Ye, Yunrui Ye, Yiting Wang, Tong Tong, Su Yao, Yao Xu, Qingru Hu, Yulin Liu, Changhong Liang, Guangyi Wang, et al. Automated assessment of necrosis tumor ratio in colorectal cancer using an artificial intelligence-based digital pathology analysis. *Medicine Advances*, 1(1):30–43, 2023.

[Yu *et al.*, 2024] Xiaotian Yu, Haoming Luo, Jiacong Hu, Xiuming Zhang, Yuexuan Wang, Wenjie Liang, Yijun Bei, Mingli Song, and Zunlei Feng. Hundredfold accelerating for pathological images diagnosis and prognosis through self-reform critical region focusing. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 1607–1615. International Joint Conferences on Artificial Intelligence Organization, 8 2024. Main Track.

[Zhao *et al.*, 2020a] Yu Zhao, Fan Yang, Yuqi Fang, Hailing Liu, Niyun Zhou, Jun Zhang, Jiarui Sun, Sen Yang, Bjoern Menze, Xinjuan Fan, et al. Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4837–4846, 2020.

[Zhao *et al.*, 2020b] Yu Zhao, Fan Yang, Yuqi Fang, Hailing Liu, Niyun Zhou, Jun Zhang, Jiarui Sun, Sen Yang, Bjoern Menze, Xinjuan Fan, et al. Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4837–4846, 2020.